Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

文化と言語の維持保存に貢献するための デジタル博物館の試み

トランスクリプションデータを流用する 字幕映像生成システムの提案

元木環[†]，上田寛人[†]，宮部誠人[†]，河原達也[†]，
林由華[††]，田窪行則[††]

映像データを Web サイト上で一般に公開するようなデジタル博物館は，様々な場所からアクセスが可能になる利点があるが，しばしば元データと展示公開データのひも付けが曖昧になること，構築後，システムのメンテナンスとコンテンツの更新を適切に行わなければ，実在の博物館よりも急速に陳腐化しがちであることが問題といえる．本研究で研究開発しているデジタル博物館は，危機言語の一つである宮古島池間の言語と文化の継承に資することを目的として，研究者が調査・記録した資料や研究成果をもとに構築・開発を進めているものである．データの管理概念を整理するとともに，研究者自身が情報やコンテンツを追加できるシステムを設計することで，資料の死蔵を防ぎ，研究の質的な向上，異分野交流の促進をはかるため，言語学研究者が日常的に作成している映像音声データからのトランスクリプションデータを流用した字幕映像を生成するシステムを提案することで，データの更新性と共同利用の可能性を促進する可能性について述べる．

The Digital Museum project for the documentation of Ikema Ryukyuan

-Proposal of subtitles-movie generation system that used transcription data

Tamaki Motoki[†] Hiroto Ueda[†] Makoto Miyabe[†]
Tatsuya Kawahara[†] Yuka Hayashi[††] Yukinori Takubo[††]

The Digital Museum Project is our attempt at the documentation of Ikema, one of the endangered dialects of Southern Ryukyuan, spoken on Miyako Islands, Okinawa, Japan. We have been studying one of the dialects of the language spoken in Nishihara, and have made recordings of natural discourse. We made the system that was able to make the subtitle for the contents of the museum as a researcher. It is sure to improve the update of data.

†京都大学学術情報メディアセンター
†† 京都大学大学院文学研究科

1. はじめに（研究の背景）

現在，琉球の言語と文化は消滅の危機に瀕している．マスメディアの発達や学校教育，特に方言札をはじめとする，標準語励行のための方言使用禁止の教育・文化政策の浸透により，その文化と母語が急速に失われかけている．言語については，ユネスコの Atlas of the World's Languages in Danger[1)]で，厳しい危機状態（severely endangered），確定的な危機状態（definitely endangered）[2) 3)]と認定されているように，琉球語を母語として自由に話せる話者はほとんどが 70 歳代後半，方言によっては 80 歳代後半でも母語として維持できていない状況である．

また，文化の継承についても，琉球はシャーマン的な儀式を最近まで残し，独自の風習，芸能を維持しながら豊かな生活を送ってきたが，生活形態の変化によって既にそれらが途絶えたり，存続が危ぶまれる状況にあり，人類学や民族学においても風習や文化に対して調査・研究がなされている．

こうした状況を憂いて，沖縄言語研究センターが設立された，1970 年代半ばには研究者や学生による方言の記録が始まっており，1980 年代後半には語彙や動詞，形容詞のみならず，ことわざやまじない，童謡，芝居の脚本等もテキストと録音記録に残す重要な仕事がなされている．[4)] 2001 年にはそれらの音声・テキストデータの一部が「琉球語音声データベース」[5)]として琉球大学附属図書館の Web サイト上で公開されているが，AV 機器や編集機器の低価格かつ高品質化，PC の普及により，現在の調査記録にはビデオなど映像が用いられることが一般化してきており，それに伴う新たな保存や公開の形態が必要と考えられる．

また，2002 年度（平成 14 年度）には，沖縄県により沖縄デジタルアーカイブ整備事業「Wonder 沖縄」が実施された結果，Web サイト上での民族博物館的デジタルアーカイブが公開されている．扱うコンテンツとして，「総ウエッブページ数 1 万ページ以上，高精細デジタル映像１０時間以上」[6)]とされ，2003 年の立ち上げ時には内容的には大規模かつ網羅的なものであり，視覚的にも最新の表現形態が用いられ工夫が凝らされていることから，その地域の文化の魅力を伝えるものとして成功しているが，以降のコンテンツの更新などメンテナンスが充分ではなく，見た目からはテキストや映像の出典や解説が必ずしも学術的な対応が取れているかどうかが分からないものである．[7)]

さらに言語学においては，1990 年代末頃より言語ドキュメンテーション研究と呼ばれる，言語の調査記述を行う調査言語学から生まれた研究領域が定義されている．中山（2009）によると，言語ドキュメンテーション研究の目的は，「個別の言語に関して，長期間にわたって幅広い範囲の活動に活用する事が出来る記録（一次データ）を体系的，包括的に蓄積することである」[8)]であり，その背景は「言語ドキュメンテーション研究は，記述研究の中でもとりわけ消滅の危機に瀕する言語（いわゆる危機言語）

Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

に関わる研究活動の中から発達してきた」とある．言語「記述」 (description) に対して言語「記録」(documentation) と呼ばれ，グローバル化にともなって世界中で多くの言語が急速に失われていっていることを背景とし，ここ 10 年ほどで急速に一次データを映像などにより「記録」することが広まっている．

欧米での最近の先行事例には， SOAS による Hans Rausing Endangered Languages Project(HRELP)[9]など，消滅の危機にある言語や文化を次世代へ残していくための大規模なデジタルアーカイブプロジェクトがあげられる．すでに多くのデータが収集，公開されており，それらは研究者だけでなく，その言語・文化の担い手である住民，一般の人たちも Web サイト上でそれらに接することができ，さらには他地域で同じ問題を抱える研究者，住民が横断して資料に触れることのできるシステム設計理念を掲げて構築している．しかし表現としては，視聴覚に訴えるような表現は見られず，調査を受けた地域住民がそれらの結果を受け取ることにおいて満足のいくものとはいえないだろう．言語学以外でも，地域研究におけるフィールドワークの問題点として，その対象者（インフォーマント）が，その調査結果を満足のいく形で受け取れなかったり，研究者がその調査の意図を伝え理解を得ることの難しさは随所で指摘がある．[10]

研究者にも調査された地域住民にとっても，いかに記録したデータを死蔵させず，有効に活用できる形で保存，公開，更新していくのかは今後の地域情報アーカイブを考える上で大きな問題であると考えられる．

2. デジタル博物館「ことばと文化—琉球列島」の紹介

2.1 プロジェクト・博物館の概要

本デジタル博物館プロジェクトは，琉球語宮古池間方言を母語とする沖縄県宮古島市西原地区において実施されているものである．現在この地区での池間方言母語話者は 400 人～600 人程度であるが，５０代でも自由に方言で話せるという，沖縄でも非常に珍しい地区である．*1 我々は，言語学調査のチームとコンテンツ作成のチームとの共同研究として他分野の研究者の協力も得ながら進めており，主に言語学調査チームが研究資料として言語や文化に関する映像や音声データを収集し，それをコンテンツ作成のチームが教材や展示コンテンツとして作成し，一次データとともにインターネット上にある博物館に収蔵，展示を繰り返すというコンセプトで構築している．コンテンツとして入りうるのは，地域住民による言語作品（方言絵本やオリジナル方言歌劇など）も含み，その発表の場としても機能する．また，言語学習のためのツールも備えており，方言を使うことのできない若い世代や，その他この方言に興味を持つ人々が学習することのできる場も用意している．実在の博物館において柱となる機能（資料収集，整理保管，調査研究，教育普及）[11]を，デジタルデータの保存管理，公開手法の場合に置き換え，わかりやすく美しい表現と更新性の両立を目指して，最終的には言語と文化の保存と維持継承に寄与することを目的に構想されている．

またこのプロジェクトは，西原地区をプロトタイプとして，琉球語，琉球文化全般に拡張し，すでに存在するデジタルデータを一箇所に集め，かつ公開できる，デジタル琉球博物館を作成するための準備としても位置づけられている．

2.2 展示による効果：メタデータの充実

本博物館で取り扱うコンテンツは，琉球諸言語の記録が中心になるが文法や音韻・形態といった言語学的関心に基づいた記述を行うとともに，

1) 映像や音声などさまざまなメディアを用いた一次データを含む包括的な言語資源を，
2) 言語学者のみならずその言語・文化に関心のある人全て，特にその言語のコミュニティ内の人が活用できるような方法で，
3) 長期保存に耐えうる形で保存する

ことが前提となっている．

研究者の一次データは，それそのものだけでは断片的な記録でしかなく，多くの人の関心を与え，意味のあるものになるためには，会話一つにとっても言葉の意味や発音だけでなく，詳しいメタデータやアノテーション，背景知識などの様々な意味づけが必要になる．ここでは，一次データはそのままの形で残しつつ，さまざまな視点による「展示」という行為を通じてコンテクストの再構築を繰り返すことにより，メタデータの充実が達成されると考えた．例えば，映像を展示することにより付けられた言語的なメタデータを文化人類学，社会学の研究者が利用できたり，文化人類学，社会学の研究者の展示の際に付した儀礼の意味に関するメタデータを言語学者が利用することができるのは，学問の質的向上につながると考えられる．また「展示」するという行為が，研究者においても，データに（他人にわかるように）メタデータをつける動機付けにも繋がると考えた．

2.3 使用技術

技術的には，XHTML+CSS で意味的に正しくコーディングするように心掛け，W3C に出来る限り準拠する形を取った．アーカイブは，将来の技術的変化に合わせてコンテンツを移動できるように考えなければならず，できる限りスタンダードな技術を使

*1西原は，池間島から明治時代に移住してできた集落であるため自己アイデンティティの確認作業を他所より行わなければならず，自分たちの母語と独自の文化を維持してきた．また，シャーマン的な願いを行う神聖な場所を維持するための組織があり，主に女性たちの手で何百年にもわたる儀式を現在にまで伝えている．しかし，この地区でさえも，他地区への移住，他地域からの流入，学校教育により，母語を失い，現代生活を続けるために伝統的な儀式を行うことが困難となっている．

Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

うことを選択している．文字情報はほとんどの場合テキストデータで表示し，映像データは FLV フォーマット，音声データは MP3 フォーマットで視聴する形を取った．どういったデータフォーマットを選択するかは，今後の技術，環境の変化により変更があると思われるが現時点ではなるべく高品質かつデータサイズが大きくないもの，特別なプラグインを用いずに，出来るだけ再生環境を選ばないという標準的な Web サイトデザインと同じ手法をとっている．

2.4 博物館の設計構造

構造としては，一般に広く公開される空間と，限定されたメンバーが立ち入ることのできる空間に分け，さらに機能別に４層の構造を構想した（図１）．

● 開架式空間（誰でもアクセスできる）

(1) 展示スペース

地域の儀礼を記録し，儀礼の世代間継承のマニュアルとして働くことや，若い世代が学べるような地域の言語の文法，辞書，会話練習帳，読本を提供すること，研究者による調査・研究成果の可視化という機能を持つ．
展示の入れ替えがこまめに可能であること．展示されるデータはオリジナルデータに辿れることが求められる．

(2) 資料室

図書館スペースや映像ライブラリとして機能する．展示スペースに使われているデータの一覧や，一般に公開可能なデータがリストアップされる．

● 閉架式空間（特定のメンバーのみアクセス可）

(3) 閉架式資料室とアーカイブ

過去の展示物をトランスクリプションとメタデータと別ファイル，かつ関連は保って格納することが求められる．
メタデータは展示によって充足されることが期待される．

(4) データ格納空間

原データが最小限のメタデータとともに格納される．

これらの層は，適切なセキュリティを設けることで，データにアクセスできる人をコントロールすると同時に，html で作成するメリットを生かし，展示データはその展示場所が異なっても複製ではなく一つのデータにリンクされるように考えられている．このことにより，メタデータも一つのオリジナルデータにひも付けされる．

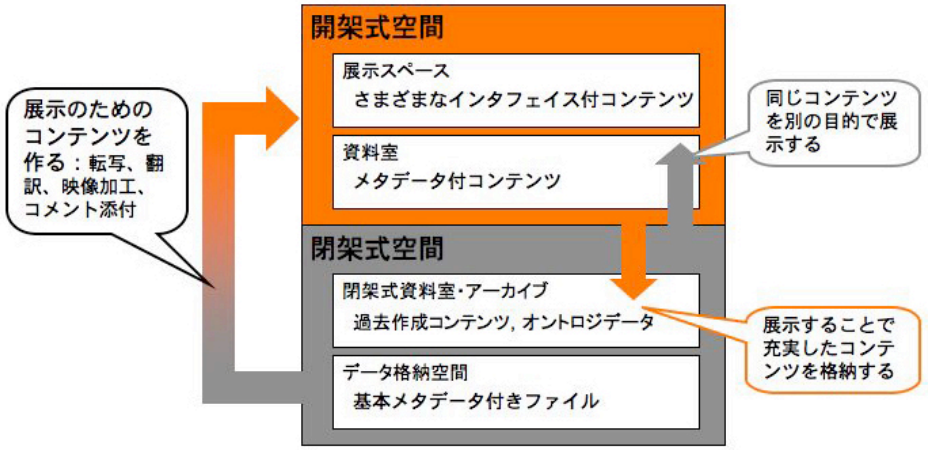

図 1　4 層の博物館の概念図

2.5 コンテンツ紹介

博物館に公開されている実際のコンテンツを紹介する．常設展のデザインは，地域が変わっても機能するよう文化人類学や民俗学の総論などに使われている項目を引用し，作成した．

主な項目は
展示室：西原の概観，西原の歴史，西原の生活，西原の一日，西原の一年（図 2）
その他小項目（一生，食，共同体，行動，信仰など）

学習室：言語の歴史（系統関係，音韻変化），日常のことば（場面と表現，文法，読本），辞書

資料室：デジタル絵本，映像番組，用語集，関連資料，アーカイブス
となっている．

Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

図 2　西原の一年（年中行事）

各項目には，解説テキストが書かれていると同時に，テキストの中のキーワードや図表の中，あるいは地図の中のあるスポットに関連する項目や辞書へのリンクがはってある．また，地域住民が方言によって語られるいくつかの映像資料を発話と訳の字幕映像付で見ることができるコンテンツが用意されているのが特徴である（図 3）．池間方言は，元は日本語と同じとはいうものの，知らないものにとってはほとんど発音を聞き取ることも不可能な初めて出会う外国語のようなものであるため，若年層やこの言葉や文化に初めて触れる人に関心を持ってもらうためにもこのような仕組みは有用である．また，調査結果を表すという点においても，撮影された結果がこのようになると説明もしやすい．

これらの映像は，プロフェッショナルではない研究者自身による撮影や，様々な人により撮影されることが想定され，画面のどこに重要なポイントをフレーミングするかはコントロール不能である．そこで映像の上の字幕で被ることを避けるため，一段の字幕を 1 つの映像として作成し，３画面同時に再生が可能なスクリプトにより再生をしている．また，言語習得の学習にも利用を想定しているため，それぞれの字幕は必要に応じて，表示／非表示を選択することができる．

図 3　字幕映像のついた映像コンテンツ

3. 字幕映像データ生成システムの提案

3.1 従来手法

前項で紹介した字幕映像付の映像コンテンツのようなページを作成するには，従来，その研究に際し作成したデータを映像編集，効果を付けるためのソフトウェアの習熟者に渡す（あるいは自らが習得して作業する）事が必要であるため，データの受け渡しにも，編集作業にも手間がかかるものであった．例えば，本博物館の場合，言語調査チームが記録した映像データと書き起こしや訳のテキストデータ，時間情報をコンテンツ作成チームが CSV などの形式で受け取り，映像編集ソフトで指定された時間情報を見ながら，手動で字幕の映像を作成し，Web サイトへアップロードするという手順になる．

しかし，研究者が会話の書き起こしをする場合，あるいは各種言語への翻訳テキストの作成，またその映像内でどういった事が行われているか注釈（annotation)を付ける場合には，通常，会話の書き起こしや分析などに特化したソフトウェアを使用するが，その際字幕に表示するテキストデータと時間情報を記述することから，このデータをもとに半自動的に字幕映像を生成できることになれば，省力化となる．また，このコンテンツは，研究者が日々大量に分析する作業の過程で生成できるため，更新性の向上も期待できる．

Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

3.2 提案手法

本提案は，会話分析やジャスチャー研究に良く使用されるソフトウェアのひとつである ELAN[*t2] で作成した字幕データを Subtitles 文書ファイル形式で書き出し，サーバ上のプログラムで，そのファイルを Subtitles 文書ファイル形式から，Flash が読み込める XML 形式に自動変換するものである．これにより，字幕テキストとタイムコードの情報を Flash の映像データに反映させて書き出すことができる．通常，ユーザは意識する必要なく，字幕に表示させたいテキストやタイミングを 1 言語（1 注釈）につき 1 ファイル作成すればよい．

次に提案手法で行うユーザ視点での作業の流れを示す．

（１）動画ファイルを準備する

1. 字幕生成用動画ファイル：ELAN に読み込み可能な形式（.mp4/.mpeg/ .mov）
2. プレイヤー再生用動画ファイル：博物館で公開する形式（.flv）

（２）字幕データを準備する

3. ELAN に字幕生成用動画ファイルを読み込み，映像を見ながら注釈（字幕テキスト），タイミングの情報をつけていく．
4. 注釈（言語）ごとにそれぞれデータを.srt 形式で保存する．

（３）素材のアップロード

作成した素材ファイルを Web サーバへアップロードする

5. 再生用動画ファイル(.flv)のアップロード
6. 言語の数分用意された（現状は 1 or 2 つ）キャプションファイル（.srt）をアップロード

（４）HTML のテンプレートを書き換えて新規ファイル作成

7. プレイヤー（JavaScript+Flash）の指定
8. タイトルの書き換え
9. キャプションファイルの指定
10. 動画ファイルのパスを指定

（５）Web サーバアップロード（完成）

11. サーバの指定ディレクトリへアップロードし，公開

*2 ELAN http://www.lat-mpi.eu/tools/elan/　各種ソフトウェア（Shoebox や Tanscriber など他のソフトウェアで記述したファイルを読み込むことが可能であり，かつアノテーションに日本語が扱えるソフトウェアであったため選択している．

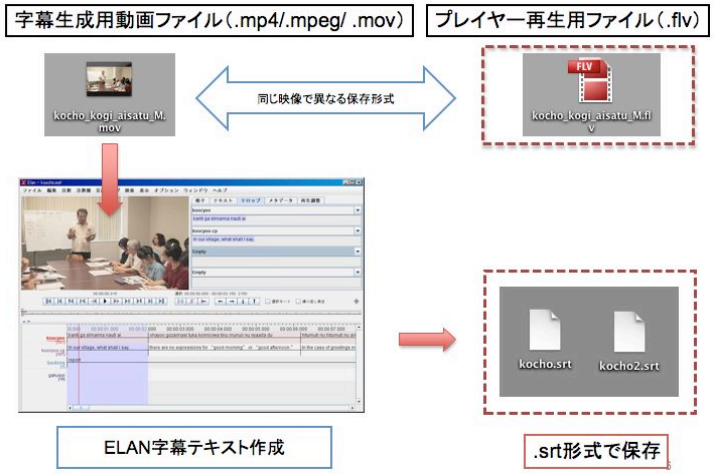

図 3　字幕映像データ生成の手順

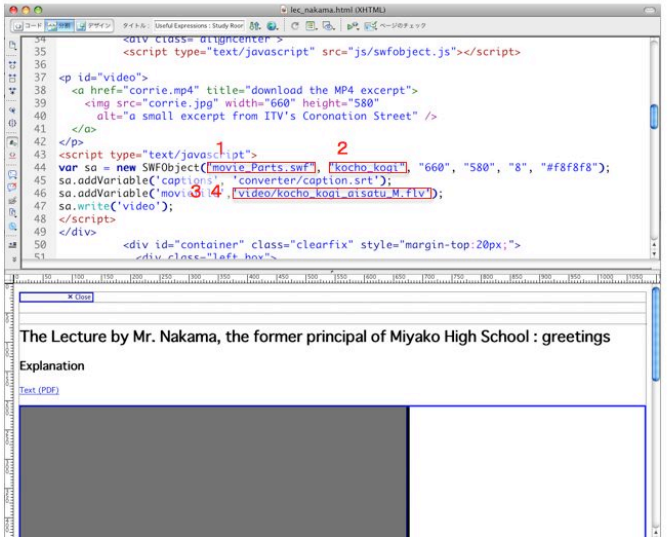

図 4　HTML テンプレートの書き換え

Vol.2010-FI-98 No.1
Vol.2010-DD-75 No.1
2010/3/4

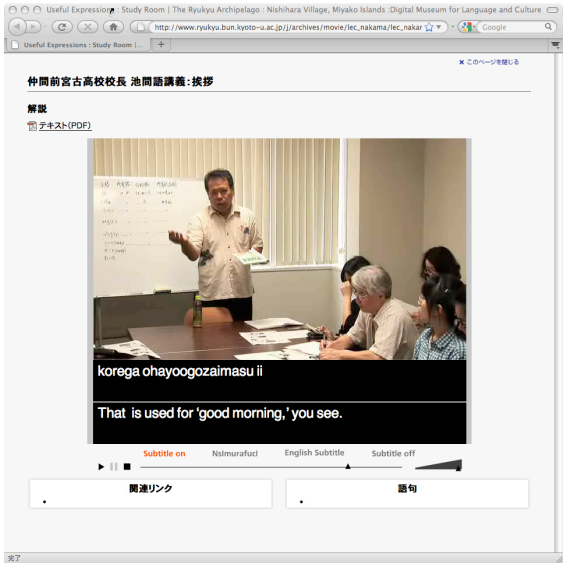

図 5　完成画面

4. まとめ

本デジタル博物館は，言語ドキュメンテーション分野の国際会議である The 1st International Conference on Language Documentation and Conservation（ICLDC）, USA, Honolulu ,2009.3 において発表され、「同じような展示スペースはどのようにすれば構築できるのか」「このようなものを自分も作ってみたいが、手伝ってくれるか」などの問い合わせをたくさんいただいた。言語学者だけではなく，アーカイビストをはじめとする技術者、記録対象となる話者コミュニティからの参加も多い中の反応で，このような「博物館展示」の機能をもつアーカイブシステムの期待と需要が確認できた．続いて作成したこの字幕映像生成システムをチーム内の言語学者で利用したところ，自身の手でコンテンツを作成することができ，コンテンツの更新性を高めるのに有用であるという認識を持つことができた．現在，近隣地区の調査・研究を行っている言語学者へ依頼し，システムの利用拡大を進めており，今後はより広い範囲から意見のフィードバックを測りつつ，コンテンツの充実と同時に字幕生成以外の更新システムの開発を進めていく予定である．

5. 謝辞

本研究は，京都大学グローバル COE プログラム「親密圏と公共圏の再編成をめざすアジア拠点」，ならびに京都大学学術情報メディアセンター「コンテンツ作成共同研究」の補助を受けて実施したものである．このプロジェクトは執筆者の他，車田千種氏，平井芽亜里氏，岩倉正司氏，高橋三紀子氏と作成を行っている．また，仲間博之氏，花城千枝子氏，仲間忠氏，赤嶺和子氏，長崎恵長氏をはじめとする地域の皆様に深く感謝いたします．

参考文献

1) UNESCO Atlas of the World's Languages in Danger
http://www.unesco.org/culture/ich/index.php?pg=00206
2) UNESCO Intangible Cultural Heritage Endangered languages
http://www.unesco.org/culture/ich/index.php?pg=00139
3) 佐々木冠：日本の言語状況ー多様性は失われるのか，月刊言語 Vol.38，No. 7，pp.8-15（2009）
4) The School of Oriental and African Studies (SOAS) ,The University of London
Hans Rausing Endangered Languages Project http://www.hrelp.org/
5) 狩俣 繁久：危機言語としての琉球方言の研究状況 日本復帰後から今日までの活動についてのおぼえがき，国立民族学博物館調査報告，No.39，pp.257-267（2003）
6) 琉球大学附属図書館「琉球語音声データベース」
http://ryukyu-lang.lib.u-ryukyu.ac.jp/
7) 沖縄デジタルアーカイブ「Wonder 沖縄」
http://www.wonder-okinawa.jp/index_jp.jsp
8) 笠羽 晴夫『デジタルアーカイブの構築と運用―ミュージアムから地域振興へ』水曜社（2004）
9) 中山俊秀：新時代の記述言語学 <上>―つながる言語記録にむけて，月刊言語 Vol.38，No. 7，pp.66-73（2009）
10) 安渓遊地『調査されるという迷惑―フィールドに出る前に読んでおく本』みずのわ出版 (2008)
11) 加藤有次『博物館機能論』雄山閣出版，pp.3-9,15-19.(2000)